基于LDA模型的餐厅推荐方法研究

张晓阳[1] 秦贵和[1,2] 邹 密[1] 孙铭会[1] 高庆洋[3]
(吉林大学计算机科学与技术学院 长春 130012)[1]
(符号计算与知识工程教育部重点实验室 长春 130012)[2] (吉林大学软件学院 长春 130012)[3]

摘 要 随着网络的飞速发展,餐饮类的评价信息数量急剧增加。对餐饮评价进行有效分析不仅能够帮助消费者进行用餐选择,还可以帮助商家对餐厅服务进行改进。为此,提出了一种基于LDA(Latent Dirichlet Allocation)模型的餐厅推荐方法。首先,对餐厅评价信息进行情感分类,获取积极评价和好评率;其次,根据LDA模型对积极评价信息文本进行聚类,生成餐厅标签;最后,计算用户需求与餐厅标签的相似度,根据相似度和好评率向用户推荐餐厅。基于通过网络获取的真实餐饮评价信息进行实验,结果表明,该方法生成的餐厅标签的效果好,能准确地向用户推荐餐厅。

关键词 评价信息,LDA,情感分析,文本聚类,餐厅标签,餐厅推荐
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.07.032

Research on Recommendation Method of Restaurant Based on LDA Model

ZHANG Xiao-yang[1] QIN Gui-he[1,2] ZOU Mi[1] SUN Ming-hui[1] GAO Qing-yang[3]
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)[1]
(Symbol Computation and Knowledge Engineer of Ministry of Education, Changchun 130012, China)[2]
(College of Software, Jilin University, Changchun 130012, China)[3]

Abstract With the rapid development of the network, the amount of the evaluation information of the food and beverage has increased dramatically. The effective analysis of the evaluation information can not only help the consumers choose the suitable restaurant, but also help the businesses improve service. For this purpose, a restaurant recommendation method based on LDA(Dirichlet Allocation Latent) model was proposed. First of all, it classifies the evaluation information according to the emotional tendencies, and then gets the positive evaluation and praise rate. Secondly, it manipulates the LDA model for text clustering to generate restaurant tags. Finally, it calculates the similarity between the user's needs and the restaurant tags, and according to the similarity and the rate of praise, recommends the suitable restaurants to customers. We got the real food and beverage comments from the Internet, and carried out the experiment. As a result, the effect of the restaurant tags produced from this method is good, which could accurately recommend the restaurants to users.

Keywords Evaluation information, LDA, Emotion analysis, Text clustering, Restaurant tags, Restaurant recommendation

1 引言

随着信息技术与网络的发展,餐饮评价信息呈现海量特征,如何在海量评价信息中获得餐厅的特征并正确地向用户推荐餐厅成为当前餐厅评价信息大数据处理的迫切需求。目前已有一些餐饮类点评网站为用户提供餐厅标签信息,例如大众点评网、美团网等。目前这类网站使用预设标签,通过用餐用户选择预设标签项,获取餐厅特征。此引导性方法获取的餐厅特征具有局限性,不能完全反映用户的客观评价。

由于餐厅的文本评价信息数量巨大、格式多样且伴随有文本噪声,使得用户难以阅读所有评价文本而形成有效知识。传统的向量空间模型VSM(Vector Space Model)[1-4]的原理是将文本分词后映射到向量空间,该模型根据词语的词频权重提取文本关键特征。常用的方法有TF-IDF[4-7]、卡方CHI(Chi-Square)[8-9]、信息增益IG(Information Gain)[10]等,但它们仅考虑词频信息并假设词之间独立,忽略了同义词的情况,同时处理海量数据时面临着维度灾难。潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)可以发现文本的潜在主

到稿日期:2016-05-20 返修日期:2016-09-22 本文受国家自然科学基金青年项目(61300145),中国博士后科学基金面上资助项目(2014M561294)资助。

张晓阳(1991—),男,硕士生,CCF会员,主要研究方向为智能控制、自然语言处理,E-mail:zhangxyjlu@163.com;秦贵和(1962—),男,教授,CCF会员,主要研究方向为智能控制、嵌入式系统、汽车电子与信息技术;邹 密(1981—),男,博士生,主要研究方向为智能交通系统、计算机视觉、模式识别;孙铭会(1983—),男,讲师,主要研究方向为车载网络安全、人机交互、物联网;高庆洋(1992—),男,硕士生,主要研究方向为智能控制、机器视觉。

题，且避免了使用 VSM 方法表示文本特征时特征向量高维稀疏的问题。文献[11-13]用 LDA 模型进行微博主题挖掘；邸亮等根据 LDA 模型设计微博用户模型，为用户推荐微博信息[14]。上述基于 LDA 模型的短文本研究都以微博文本为研究对象。与微博文本不同，餐厅评价信息不是针对某一事件或事物的简单陈述，情感信息更明显，若直接对评价信息运用 LDA 模型提取主题下的关键词，则并不能区分其情感特征，也不能向用户准确地推荐餐厅。

基于此，本文提出了一种基于 LDA 模型的餐厅推荐方法。主要从以下几个方面进行研究：获取网上真实的评价数据；根据自定义词典对评价信息精确分词；设计朴素贝叶斯文本分类器，将评价信息自动分为好评与差评两类；运用 LDA 模型生成餐厅标签；根据用户需求与餐厅标签的相似度及餐厅好评率向用户推荐餐厅；餐厅也可根据生成的标签做相应改进，不断提高服务质量。

2　LDA 模型

2.1　LDA 模型思想

LDA 是由 Blei 等人于 2003 年提出的概率增长模型，是包含文档、主题和词的三层贝叶斯模型[15]。LDA 为无监督的学习模型，不需要训练大量数据，适合处理大规模文档集合，且可以在文档中搜索出隐含的主题分布信息。相比于关键词匹配技术，LDA 主题模型的优点在于：LDA 模型更关注文档的语义信息，是一种抽象层次更高的匹配技术。模型的符号及其说明如表 1 所列。

表 1　LDA 模型符号说明

符号	说明
K	主题数量
M	文档数量
V	词语总数
N_m	文档 m 中词语的数量
$Z_{m,n}$	文档 m 中的第 n 个词的主题
$W_{m,n}$	文档 m 中的第 n 个词语
$\vec{\alpha}$	每个 m 下 Topic 的多项分布的 Dirichlet 先验参数
$\vec{\beta}$	每个 Topic 下词的多项分布的 Dirichlet 先验参数
$\vec{\vartheta}_m$	文档 m 的主题分布，$\theta=\{\vec{\vartheta}_m\}_{m=1}^{M}$ ($M\times K$ matrix)
$\vec{\varphi}_k$	主题 k 下的词语分布，$\Phi=\{\vec{\varphi}_k\}_{k=1}^{K}$ ($K\times V$ matrix)

对于整个文档集合 M，LDA 假设其中有 K 个独立的主题；对于每一个文档 m，认为其是由 K 个主题随机组成，且每个文档 m 对于 K 个主题呈多项式分布，该多项式概率分布中文档的先验概率分布满足 Dirichlet 分布；每个主题 topic 是词语上的多项式分布，且每一个主题 topic 中词的先验概率分布是 Dirichlet 分布。

将 LDA 看作一个生成过程，描述如下：

For each topic $k\in[1,K]$:

　　Draw a multinomial $\vec{\varphi}_k$ from a Dirichlet prior $\vec{\beta}$;

For each document $m\in[1,M]$:

　　Draw a multinomial $\vec{\vartheta}_m$ from a Dirichlet prior $\vec{\alpha}$;

For each word $n\in[1,N_m]$ in document m:

　　Draw a topic $z_{m,n}$ from multinomial $\vec{\vartheta}_m$;

　　Draw a word $w_{m,n}$ from multinomial $\vec{\varphi}_{z_{m,n}}$;

其图形表述如图 1 所示。

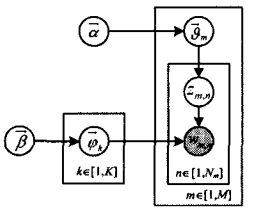

图 1　LDA 主题模型图形表示

图 1 中变量的类型有两种，其中潜在变量用空心圆来表示，可观测变量用实心圆来表示，两变量间的条件依赖用箭头标明，方框代表重复抽样。对于一个文档集合 M，α 和 β 是先验参数，根据经验值给出；w 为已知观测变量。根据图 1 得出整个模型所有参数的联合分布概率：

$$P(\vec{w}_m,\vec{z}_m,\vec{\vartheta}_m,\Phi)=\prod_{n=1}^{N_m}P(w_{m,n}|\vec{\varphi}_{z_{m,n}})P(z_{m,n}|\vec{\vartheta}_m)P(\vec{\vartheta}_m|\vec{\alpha})P(\Phi|\vec{\beta}) \quad (1)$$

2.2　Gibbs Sampling

LDA 是一个相对简单的模型，但由于存在多个未知变量，且变量之间存在耦合现象难以进行精确推理，因此通常使用 Gibbs 抽样进行近似推理。Gibbs 抽样是马尔可夫链蒙特卡尔理论(Markov Chain Monte Carlo，MCMC)中用来获取一系列近似等于指定多维概率分布观察样本的算法[16]。

式(1)中，主题的多项式分布 θ 由 α 生成，词语多项式分布 Φ 由 β 生成。因此式(1)等价于式(2)的联合概率分布。

$$P(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta})=P(\vec{w}|\vec{z},\vec{\beta})P(\vec{z}|\vec{\alpha}) \quad (2)$$

式(2)的求解分为两个部分，第一个部分 $P(\vec{w}|\vec{z},\vec{\beta})$ 是根据主题 $\vec{z}$ 和先验参数 β 采样词语的过程，第二部分 $P(\vec{z}|\vec{\alpha})$ 是根据先验参数 α 采样主题的过程。

第一个因子 $P(\vec{w}|\vec{z},\vec{\beta})$ 根据确定的主题 $\vec{z}$ 和由先验参数 β 采样生成的多项式分布 Φ 得到：

$$P(\vec{w}|\vec{z},\vec{\beta})=\int P(\vec{w},|\vec{z},\Phi)P(\Phi|\vec{\beta})\mathrm{d}\Phi$$

$$=\prod_{z=1}^{K}\frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})},\vec{n}_z=\{n_z^{(v)}\}_{v=1}^{V} \quad (3)$$

其中，$n_z^{(v)}$ 表示词语 v 在主题 z 中出现的次数。式(3)是 K 个 Dirichlet-Multinomial 模型的乘积。同理，根据式(3)的推导过程得到第二个因子的公式展开：

$$P(\vec{z}|\vec{\alpha})=\int P(\vec{z}|\theta)P(\theta|\vec{\alpha})\mathrm{d}\theta$$

$$=\prod_{m=1}^{M}\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})},\vec{n}_m=\{n_m^{(k)}\}_{k=1}^{K} \quad (4)$$

其中，$n_m^{(k)}$ 代表主题 k 在文章 m 中出现的次数。综合式(3)、式(4)，得到 $P(\vec{w},\vec{z})$ 的联合分布结果：

$$P(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta})=\prod_{z=1}^{K}\frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})}\cdot\prod_{m=1}^{M}\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})} \quad (5)$$

通过联合分布 $P(\vec{w},\vec{z})$ 计算给定观测变量 $\vec{w}$ 下的隐变量 z 的条件分布：

$$P(z_i=k|\vec{z}_{\neg i},\vec{w})\propto P(z_i=k,w_i=v|\vec{z}_{\neg i},\vec{w}_{\neg i})$$

$$=E(\vartheta_{m,k})\cdot E(\varphi_{k,v})=\hat{\vartheta}_{m,k}\cdot\hat{\varphi}_{k,v} \quad (6)$$

根据狄利克雷分布的特性求解 Dirichlet 分布期望：

$$\vartheta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_m^{(k)}+\alpha_k)},\varphi_{k,v}=\frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)} \tag{7}$$

将式(7)代入式(6)得：

$$P(z_i=k|\vec{z}_{\neg i},\vec{w})\propto\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(k)}+\alpha_k)}\cdot\frac{n_{k,\neg i}^{(v)}+\beta_v}{\sum_{v=1}^{V}(n_{k,\neg i}^{(v)}+\beta_v)} \tag{8}$$

式(8)对应为 $P(\text{topic}|\text{doc})\cdot P(\text{word}|\text{topic})$，这个概率值对应着 doc-topic-word 的路径概率。因此，K 个 topic 对应着 K 条路径，Gibbs Sampling 在这 K 条路径中进行采样，如图 2 所示。

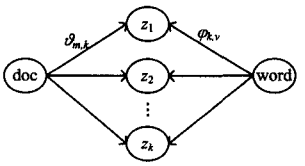

图 2 Gibbs 采样路径

3 餐厅推荐模型

3.1 关键特征提取

餐饮评价信息为用户对餐厅环境、卫生以及菜品种类、口感等方面的主观评价，其语言描述基本不具有规范性，且评价文本也相对较短。为了凑字数，部分用户会随意书写一些与餐饮无关的标点符号等信息，这样的评价信息为噪声。本文首先去掉评价文本中的打分、用户 ID 等信息，保留对餐饮情况的评价信息，再根据评价信息的长度进行过滤，删除少于 20 字的评价。然后，使用 jieba 分词工具对评价信息进行分词。将收集整理的八百多道菜名和几十家饭店的名称以及地理位置添加到自定义词典中，实现评价信息中菜名、饭店名、街道名称等的正确切分。对于评价信息，一般的消极评价为否定形式，如“梅菜扣肉很不好吃”，对其进行分词的结果为：梅菜扣肉 n/很 d/不 d/好吃 v/，分词之后将“不”和“好吃”分开，若将词语作为文本信息的特征，则会出现歧义。基于此，本文将这类词语进行组合形成新的形容词短语，并将其加入自定义词典中，这样可消除大部分歧义问题。进行精确分词后根据停用词表和词性去掉停用词，可降低文本信息的维度，同时根据同义词词典对同义词做归一化处理，将出现的同义词按照词典的映射关系转化为某一特定的词。关键特征提取的过程如图 3 所示。

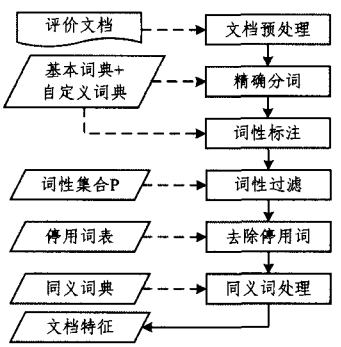

图 3 关键特征的提取过程

3.2 评价信息情感分析

用户对餐饮的评价与微博不同：微博主要是对事件或事物的描述；而餐饮评价内容的情感信息更突出，通常可分为好评和差评两类。对大量评价信息进行人工分类是很耗时的，因此构建贝叶斯文本分类器来对评价信息进行分类。

3.2.1 Naive Bayes 的基础知识

贝叶斯全概率公式为：

$$P(c_i|w_1,w_2,\cdots,w_n)=\frac{P(w_1,w_2,\cdots,w_n|c_i)P(c_i)}{P(w_1,w_2,\cdots,w_n)} \tag{9}$$

其中，c 表示文本的类别集合，$P(c_i)$ 为类别 i 出现的概率，$P(w_1,w_2,\cdots,w_n|c_i)$ 为特征向量 $(w_1,w_2,\cdots,w_n)$ 在文本类别为 c_i 时出现的概率，$P(w_1,w_2,\cdots,w_n)$ 为特征向量出现的概率。假设特征在文本中出现的概率是独立的，即词和词之间不相关，则联合概率可以表示为乘积的形式：

$$P(c_i|w_1,w_2,\cdots,w_n)=\frac{P(w_1|c_i)P(w_2|c_i),P(w_n|c_i)P(c_i)}{P(w_1)P(w_2)\cdots P(w_n)} \tag{10}$$

3.2.2 朴素贝叶斯的两种模型

(1)多项式模型(multinomial model)——词频型

在多项式模型中，设某文档 $d=(v_1,v_2,\cdots,v_n)$，v_i 是文档 d 的关键特征词。

先验概率为：

$$P(c)=\frac{N_c}{N} \tag{11}$$

其中，N_c 表示类别 c 下的所有单词数量，N 表示训练样本中所有单词的数量。

类条件概率为：

$$P(v|c)=\frac{N_{c,v}+1}{N_c+|T|} \tag{12}$$

其中，$N_{c,v}$ 是在类别 c 下单词 v 在每个文档中出现的次数和；T 代表训练样本中所有单词的数量(重复单词计算一次)，$|T|$ 为训练样本中单词的种类数量。$P(v|c)$ 代表单词 v 为文本 d 划分为类别 c 所提供证据的概率。

(2)伯努利模型(Bernoulli model)——文档型

该模型的粒度较大，以文件为粒度，采用二项分布模型，只考虑事件发生或者不发生，每个单词的表示变量是布尔型的类条件概率。

$$P(c)=\frac{M_c}{M} \tag{13}$$

$$P(v|c)=\frac{M_{c,v}+1}{M_c+2} \tag{14}$$

其中，M_c 是被标记为类别 c 的文件总数，M 是整个训练样本的文件总数；$M_{c,v}$ 是类别 c 下含有单词 v 的文件总数。

3.3 餐厅推荐算法

根据 LDA 主题-词概率分布，提取出每个餐厅主题下的前 N 个词作为该餐厅的标签，所有餐厅标签形成文档集合 $D(d_1,d_2,\cdots,d_n)$，其中 d_i 为餐厅 i 的标签集合形成的文档。用户对餐饮的需求如口味、环境等提出要求 Q，根据 BM25[17] 公式计算用户需求和每个餐厅标签 d 的相似度，相似度计算公式为：

$$Score(Q,d)=\sum_{i=1}^{n} IDF(q_i)\frac{f_i(k_1+1)}{f_i+k_1(1-b+b\frac{d_l}{avgd_l})} \quad (15)$$

其中，q_i 是对用户需求 Q 进行特征提取后的某个特征词；d_l 是文档 d 的长度，$avgd_l$ 是所有文档的平均长度；k_1 和 b 为调节因子，b 越大对文档长度的惩罚越大，k_1 用于调整词频，k_1 越大则词频的作用越大，一般根据经验设置 $k_1=1.2$，$b=0.75$[17-18]；f_i 为关键特征 q_i 在 d 中的频率；$IDF(q_i)$ 的计算公式为：

$$IDF(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \quad (16)$$

其中，N 取值为待检索的全部文档的数量，$n(q_i)$ 为 N 中包含 q_i 的文档数。

根据式(15)、式(16)得到每个餐厅标签与用户需求的相似度，根据相似度对餐厅进行排序，若相似度相同，则根据朴素贝叶斯文本分类统计出的好评率进行排序。

3.4　餐厅推荐系统

根据上述研究，设计餐厅推荐系统，其主要由 5 个部分组成。

(1)数据采集：收集并整理餐厅评价数据，进行预处理，包括删除用户 ID、打分、较短或重复评价。

(2)获取关键特征：对评价信息进行中文分词，对分词后的文本进行词性过滤、去停用词、同义词归一化处理等操作后得到评价信息的关键特征。

(3)情感分析：根据朴素贝叶斯分类器，将评价信息分为好评和差评两类；统计每个餐厅的好评率。

(4)生成餐厅标签：根据 LDA 模型生成餐厅-词分布概率，选取前 N 个关键词作为餐厅标签。

(5)餐厅推荐：根据 BM25 计算公式打分，选取与用户需求最相关的餐厅并显示结果。

系统结构如图 4 所示。

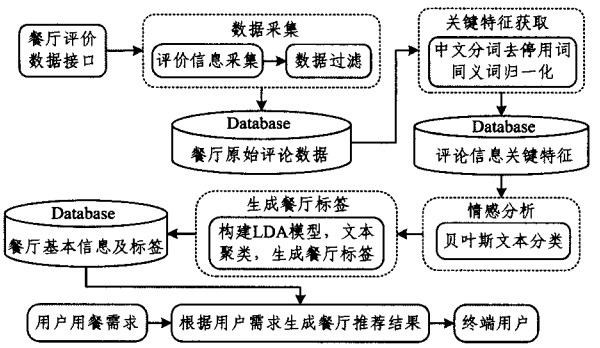

图 4　餐厅推荐系统结构图

4　实验

4.1　实验环境及数据

本文在 Windows7 64 位操作系统下采用 python 语言在 PyCharm 编程环境中编写测试程序。

通过网络获取大众点评网上 4 家长春市餐饮店的评价信息各 4000 条，每条评价信息作为一个文本文档。通过本文方法获取每个评价文本的特征向量，其具体数据信息如表 2 所列。

表 2　评价数据

餐厅名称	评价数据	特征词数量
三俞竹苑	4000	39580
东方饺子王	4000	39424
春园烤肉	4000	37810
元盛居	4000	36646
总计	16000	153460

4.2　实验结果及分析

首先，将 4 家餐厅的评价数据按3:1的比例分成两个部分：训练集和测试集，人工将训练集分为好评和差评两个子集，并为每条评论加上对应的标签；然后，训练文本分类器；最后，统计测试集中每个餐厅的好评率。好评率计算公式为：好评率＝测试集中分类为好评的文本数/测试集文本总数。结果如表 3 所列。

表 3　好评率统计表

店名	训练数据	测试数据	好评率/%
三俞竹苑	3000	1000	88.5
东方饺子王	3000	1000	83.4
春园烤肉	3000	1000	84.7
元盛居	3000	1000	85.2

对 4 家餐厅的好评信息进行 LDA 模型训练，设置参数 $\alpha=0.1$，$\beta=0.01$，此为经验值，多次实验表明，该值在本文实验数据上有较好表现。抽样主题个数 $K=4$，迭代次数为 1000。根据得到的主题-文档概率分布对文本进行聚类，通过计算查全率、查准率和 F1 值来评价该模型的聚类效果。

$$\text{查准率}(Precision)=\frac{Count(Correct\_Topic(i))}{Count(Topic(i))} \quad (17)$$

$$\text{查全率}(Recall)=\frac{Count(Correct\_Topic(i))}{Num(Topic(i))} \quad (18)$$

其中，$Count(Correct\_Topic(i))$ 为正确分类为主题 $Topic(i)$ 的评价信息数量，$Count(Topic(i))$ 为分类为 $Topic(i)$ 的评价信息数量，$Num(Topic(i))$ 为主题为 $Topic(i)$ 的评价信息总测试数量。统计结果如表 4 所列。

表 4　聚类结果(查全率、查准率和 F1 值)/%

餐厅名称	Precision	Recall	F-measure
三俞竹苑	89.95	85.00	87.40
东方饺子王	82.73	86.70	84.67
春园烤肉	84.16	79.70	81.87
元盛居	84.62	89.70	87.09

由表 4 可以看出，4 家餐厅评价数据的平均查准率为 85.37%，平均查全率为 85.28%，聚类的效果较好。为了评估本文实现的基于 LDA 模型的方法对文本分类的性能，通过实验对本文方法与传统 LDA 模型和 SVM 文本分类算法进行比较，SVM 分类器使用 libsvm。实验结果如图 5 所示。

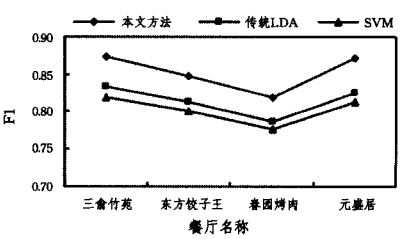

图 5　3 种方法 F1 值的对比

由图 5 可知，本文实现的方法由于对文本进行了精确分词，同时去除了同义词噪声干扰，因此文本分类的 F1 值明显高于传统 LDA 模型和 SVM 方法。相较于传统 LDA 模型，本方法的平均 F1 值由 81.38% 提高到 85.26%，提高了 3.88%；相较于 SVM，本方法的平均 F1 值由 80.18% 提高到 85.26%，提高了 5.08%。实验结果表明，本文实现的方法具有较好的分类性能。

表 5 列出了主题-词概率分布结果。可以看出，对评价文档处理后再进行 LDA 主题特征提取的主题分类效果较好。Topic1 是对“元盛居”的评价，可以看出这是一家老字号的炭火火锅店，羊肉是最受欢迎的火锅食材。Topic2 是对“东方饺子馆”的评价，可以看出其服务好，且猪肉、虾仁馅的饺子味道不错。Topic3 是对“三俞竹苑”的评价，可以看出这是一家川菜馆，其招牌菜是水煮鱼。Topic4 是对“春园烤肉”的评价，可以看出这是一家自助餐厅，环境、菜品都不错，但是存在经常排队的情况。

表 5 LDA-Gibbs 模型的主题-词与频率

Topic1		Topic2		Topic3		Topic4	
环境	0.0391	饺子	0.0780	水煮鱼	0.0436	不错	0.0359
好吃	0.0367	不错	0.0346	不错	0.0398	烤肉	0.0264
羊肉	0.0360	好吃	0.0282	好吃	0.0385	好吃	0.0224
味道	0.0341	味道	0.0255	味道	0.0334	自助	0.0220
火锅	0.0295	喜欢	0.0224	喜欢	0.0247	喜欢	0.0187
不错	0.0201	服务	0.0147	川菜	0.0210	排队	0.0169
麻酱	0.0186	吃饺子	0.0137	排骨	0.0200	春园	0.0151
感觉	0.0177	东方饺子王	0.0135	环境	0.0170	种类	0.0131
喜欢	0.0171	环境	0.0123	官保鸡丁	0.0097	环境	0.0123
炭火	0.0152	小菜	0.0112	推荐	0.0094	菜品	0.0118
朋友	0.0134	团购	0.0087	三俞竹苑	0.0083	烤鸭	0.0091
服务	0.0120	猪肉	0.0069	服务	0.0066	蛋糕	0.0087
火锅店	0.0091	水饺	0.0069	排队	0.0056	水果	0.0075
老字号	0.0084	虾仁	0.0068	干煸四季豆	0.0055	点评	0.0074
元盛居	0.0075	价格	0.0056	价格	0.0054	朋友	0.0072
…	…	…	…	…	…	…	…

根据表 5 结果，每个主题 Topic 对应一家餐厅，取每个餐厅中前 N 个词语作为该餐厅的标签。用户根据需求查找与之最相似的餐厅信息；餐厅通过对比其它餐厅的标签发现自己的不足，完善餐厅服务，提高用餐质量。

结束语 本文通过对餐饮评价信息文本进行分析，实现了一种餐厅推荐方法。将评价信息进行情感分析，获取积极评价后再结合 LDA 模型提取每个餐厅的标签。将海量评价数据转化为几个或十几个关键词语，不仅能方便用户浏览查询，还可以根据用户需求进行智能推荐。实验表明，该方法优于传统 LDA 模型及 SVM 方法，具有一定的实用价值。在今后的研究工作中，将继续优化该方法的效率。

参 考 文 献

[1] ZHOU X G, GAO F, SUN Y. Sub-topic detection and tracking based on dependency connection weights for vector space model [J]. Journal on Communications, 2013, 34(8): 1-9. (in Chinese)
周学广，高飞，孙艳. 基于依存连接权 VSM 的子话题检测与跟踪方法[J]. 通信学报，2013，34(8)：1-9.

[2] TURNEY P D, PANTEL P. From frequency to meaning: vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2015, 37(4): 141-188.

[3] LIN Y S, LIANG L, CUI Y, et al. Intelligent medical guide system based on vsm weight imporvement algorithm[J]. Computer Applications and Software, 2015, 32(9): 81-83, 111. (in Chinese)
林予松，梁璐，崔勇，等. 基于 VSM 权重改进算法的智能导医系统[J]. 计算机应用与软件，2015，32(9)：81-83，111.

[4] HUANG C H, YIN J, HOU F. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method [J]. Chinese Journal of Computers, 2011, 34(5): 856-864. (in Chinese)
黄承慧，印鉴，侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报，2011，34(5)：856-864.

[5] HOGENBOOM A, BAL D, FRASINCAR F, et al. Exploiting emoticons in polarity classification of text[J]. Journal of Web Engineering, 2015, 14(1): 22-40.

[6] ZHOU Y, DAI M H. News Recommendation Technology Combining Semantic Analysis with TF-IDF Method[J]. Computers Science, 2013, 40(S2): 267-269, 300. (in Chinese)
周由，戴牡红. 语义分析与 TF-IDF 方法相结合的新闻推荐技术[J]. 计算机科学，2013，40(S2)：267-269，300.

[7] RAMAGE D, HEYMANN P, MANNING C D, et al. Clustering the tagged web[C]// International Conference on Web Search & Web Data Mining. 2009: 54-63.

[8] PEI Y B, LIU X X. Study on improved CHI for feature selection in Chinese text categorization. Computer Engineering and Applications[J]. Computer Engineering and Applications, 2011, 47(4): 128-130, 194. (in Chinese)
裴英博，刘晓霞. 文本分类中改进型 CHI 特征选择方法的研究[J]. 计算机工程与应用，2011，47(4)：128-130，194.

[9] QIU Y F, WANG W, LIU D Y, et al. CHI feature selection method based on variance[J]. Application Research of Computers, 2012, 29(4): 1304-1306. (in Chinese)
邱云飞，王威，刘大有，等. 基于方差的 CHI 特征选择方法[J]. 计算机应用研究，2012，29(4)：1304-1306.

[10] REN Y G, YANG R J, YIN M F, et al. Information-gain-based Text Feature Selection Method[J]. Computer Science, 2012, 39(11): 127-130. (in Chinese)
任永功，杨荣杰，尹明飞，等. 基于信息增益的文本特征选择方法[J]. 计算机科学，2012，39(11)：127-130.

[11] ZHAO F, ZHU Y, JIN H, et al. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment[J]. Future Generation Computer Systems, 2016, 65: 196-206.

[12] WANG L R, YU Z T, WANG Y B, et al. Micro-blogging topic mining based on supervised LDA user interest model[J]. Journal of Shandong University (Natural Science), 2015, 50(9): 36-41. (in Chinese)
王立人，余正涛，王炎冰，等. 基于有指导 LDA 用户兴趣模型的微博主题挖掘[J]. 山东大学学报(理学版)，2015，50(9)：36-41.

(下转第 214 页)

结束语 针对 K-NN 算法在大数据环境下的不足，提出了一种基于哈希技术和 MapReduce 的大数据集 K-近邻算法。与同类算法相比，所提算法具有如下两个特点：1)算法思想简单，易于编程实现；2)算法运行效率高，在保持分类能力的前提下，所消耗的运行时间远远少于 MR-K-NN 所消耗的运行时间。未来的工作包括：1)在更多、更大的数据集上进行实验，并对实验结果进行统计分析；2)对 SimHash 算法做进一步改进，以提高其计算精度。

参 考 文 献

[1] WU X D, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining [J]. Knowledge & Information Systems, 2007, 14 (1): 1-37.

[2] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[3] ZHANG X G. Pattern Recognition(Third Edition)[M]. Beijing: Tsinghua University Press, 2010. (in Chinese)
张学工. 模式识别(第 3 版) [M]. 北京：清华大学出版社，2010.

[4] 王晓东. 计算机算法设计与分析(第 4 版)[M]. 北京：电子工业出版社，2012.

[5] ALSUWAIYEL M H. Algorithms Design Techniques and Analysis(English Edition, Second Edition)[M]. Beijing: Publishing House of Electronics Industry, 2013. (in Chinese)
ALSUWAIYEL M H. 算法设计与分析(英文版，第 2 版)[M]. 北京：电子工业出版社，2013.

[6] CORMEN T H, LEISERON C E, RIVEST R L, et al. Introduction to Algorithms(English Edition, Second Edition)[M]. Beijing: Higher Education Press, 2001. (in Chinese)
CORMEN T H, LEISERON C E, RIVEST R L, 等. 算法导论(英文版，第 2 版)[M]. 北京：高等教育出版社，2001.

[7] HART P E. The condensed nearest neighbor rule[J]. IEEE Transaction on Information Theory, 1968, 14(5): 515-516.

[8] WILSON D R, MARTINEZ T R. Reduction techniques for instance-based learning algorithms[J]. Machine Learning, 2000, 38(3): 257-286.

[9] BRIGHTON C, MELLISH C. Advances in instance selection for instance-based learning algorithm[J]. Data Mining and Knowledge Discovery, 2002, 6(2): 153-172.

[10] OLVERA-LÓPEZ J A, CARRASCO-OCHOA J A, MARTÍNEZ-TRINIDAD J F, et al. A review of instance selection methods[J]. Artificial Intelligence Review, 2010, 34(2): 133-143.

[11] BENTLEY J L. Multidimensional Binary Search Trees Used for Associative Searching[J]. Communications of the Acm, 1975, 18 (9): 509-517.

[12] OMOHUNDRO S M. Efficient algorithms with neural network behavior[J]. Journal of Complex Systems, 1987, 1(2): 273-347.

[13] UHLMANN J K. Satisfying general proximity / similarity queries with metric trees[J]. Information Processing Letters, 1991, 40(4): 175-179.

[14] KNUTH D. Art of Computer Programming, Volume 3: Sorting and searching[M]. New Jersey: Addison-Wesley Professional, 1997.

[15] GIONIS A, INDYK P, MOTWANI R. Similarity search in high dimensions via hashing [C]// Proc. of 25th International Conference on Very Large Data Bases. 1999: 518-529.

[16] SALAKHUTDINOV R, HINTON G. Semantic hashing [J]. International Journal of Approximate Reasoning, 2009, 50 (7): 969-978.

[17] JUN W, SANJIV K, SHIH F C. Semi-supervised hashing for large-scale search [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(12): 2393-2406.

[18] LI W J, ZHOU Z H. Learning to hash for big data: Current status and future trends [J]. Chinese Science Bulletin, 2015, 60(5/6): 485-490. (in Chinese)
李武军，周志华. 大数据哈希学习：现状与趋势[J]. 科学通报，2015, 60(5/6): 485-490.

[19] WANG J, LIU W, KUMAR S, et al. Learning to Hash for Indexing Big Data-A Survey [J]. Proceedings of the IEEE, 2016, 104 (1): 34-57.

[20] MANKU G S, JAIN A, SARMA A D. Detecting near-duplicates for web crawling [C]// International Conference on World Wide Web. ACM, 2007: 141-150.

[21] DEAN J, GHEMAWAT S. MapReduce: Simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.

[22] 黄宜华，苗凯翔. 深入理解大数据-大数据处理与编程实践[M]. 北京：机械工业出版社，2014.

(上接第 184 页)

[13] CHEN W T, ZHANG X M, LI Z J. Analysis of Topic Models on Modeling MicroBlog User Interestingness[J]. Computer Science, 2013, 40(4): 127-130, 135. (in Chinese)
陈文涛，张小明，李舟军. 构建微博用户兴趣模型的主题模型的分析[J]. 计算机科学，2013, 40(4): 127-130, 135.

[14] DI L, DU Y P. Application of LDA Model in Microblog User Recommendation[J]. Computer Engineering, 2014, 40(5): 1-6, 11. (in Chinese)
邸亮，杜永萍. LDA 模型在微博用户推荐中的应用[J]. 计算机工程，2014, 40(5): 1-6, 11.

[15] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[16] FOX C, PARKER A. Convergence in variance of Chebyshev accelerated Gibbs samplers[J]. Siam Journal on Scientific Computing, 2014, 36(1): A124-A147.

[17] ROBERTSON S E, WALKER S, JONES S, et al. Okapi at TREC-3[C] // Proceedings of the Third Text Retrieval Conference(TRCE 1994). Gaithersburg, USA, 1994.

[18] SHAO K, ZHANG J W. Research on personalized recommendation of Web text mining based on BM25F model[J]. Information Studies: Theory & Application, 2013, 36(11): 118-122. (in Chinese)
邵康，张建伟. 基于 BM25F 模型的 Web 文本挖掘个性化推荐研究[J]. 情报理论与实践，2013, 36(11): 118-122.